Current Approaches to 3-D Sound Reproduction

Human Factors Research Symposium 2004
NASA Ames Research Center
Moffett Field, CA, October 18-21

Elizabeth M. Wenzel
Spatial Auditory Displays Lab
Human Factors Research & Technology Division
NASA Ames Research Center
bwenzel@mail.arc.nasa.gov

Talk Overview

Ames Research Center

- Human Performance Advantages of Spatial Sound
- Spatial Cues & Perceptual Errors
- Perceptual Research & Engineering Compromises
- Dynamic VAEs: Impact of Latency
- Comparison of some current VAE Systems
- NASA's SLAB System:
  Software developed by Joel Miller
- Conclusions

VIRTUAL ACOUSTIC ENVIRONMENTS
Performance Advantages of 3D Sound

Ames Research Center

Enhanced Situational Awareness

- Direct representation of spatial information.
- Omnidirectional Monitoring: "The function of the ears is to point the eyes."
- Reinforces (or replaces) information in other modalities.
- Enhances the sense of presence or realism.

Enhanced Multiple Channel Presentation

- The "Cocktail Party Effect:" Improves intelligibility, discrimination & selective attention among voices in a background of noise or other voices ("natural" noise cancellation.)
- Enhanced Stream Segregation: Allows separation of multiple sounds into distinct "objects."

Applications of Spatial Sound

Aeronautics: ATC Displays, Cockpit Warning Systems

- Direct representation: incoming aircraft locations; left vs. right engine malfunctions
- Symbolic representation: different aircraft systems mapped to different locations (a cockpit data space)
- Enhanced intelligibility / separation: simultaneous radio communications

Telerobotic Control: Space Station Construction and Repair

- Direct representation: contact cues; range-finding
- Enhanced intelligibility / separation: simultaneous icons / symbologies

Applications of Spatial Sound

Architectural Acoustics: Acoustical CAD/CAM Systems

- Direct representation of spatial information: "auralization" of rooms
- Enhanced source intelligibility / separation: simultaneous sources

Data Spaces: Large-Scale Databases / Information Systems

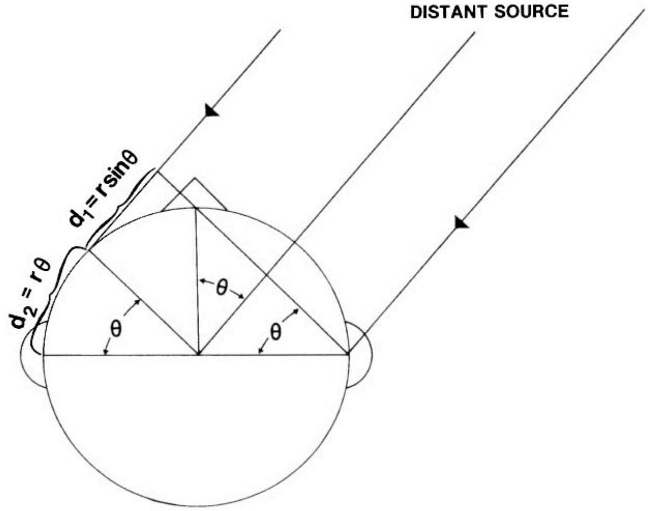
- Symbolic representation via spatial location: database navigation, architecural / spatial metaphor for database organization
- Enhanced intelligibility / separation: simultaneous icons / symbologies

Data Visualization: Computational Fluid Dynamics, Virtual Wind Tunnel

- Direct representation: airflow / noise patterns produced by aircraft engines
- Symbolic representation: localized intensities => size of error measures dispersed over the sample grid of a flow model
- Enhanced intelligibility / separation: simultaneous icons / symbologies

Spatial Cues & Perceptual Errors

Lateral localization of auditory images



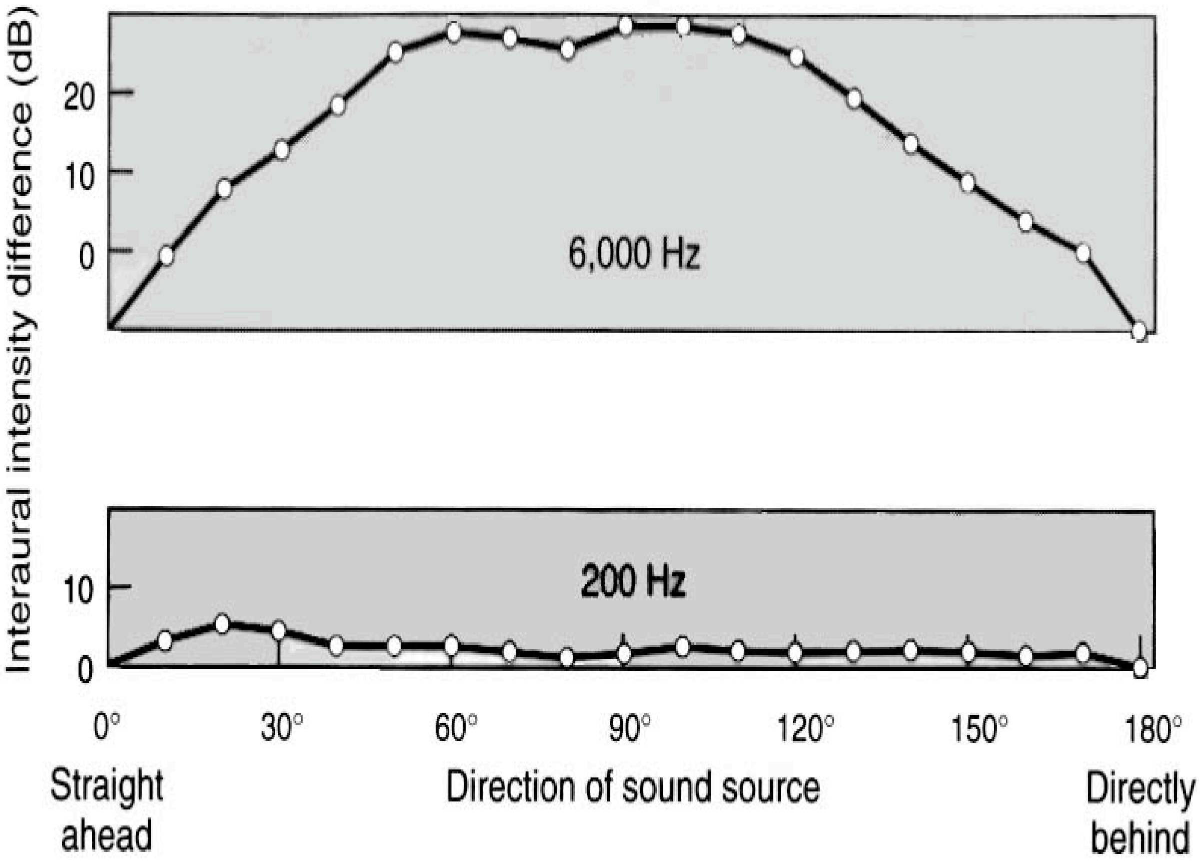
“Duplex” theory of localization

- ILD (interaural level difference): Sources off to one side are louder at the near ear due to head-shadowing.

- ITD (interaural time difference): Sources off to one side arrive sooner at the near ear.

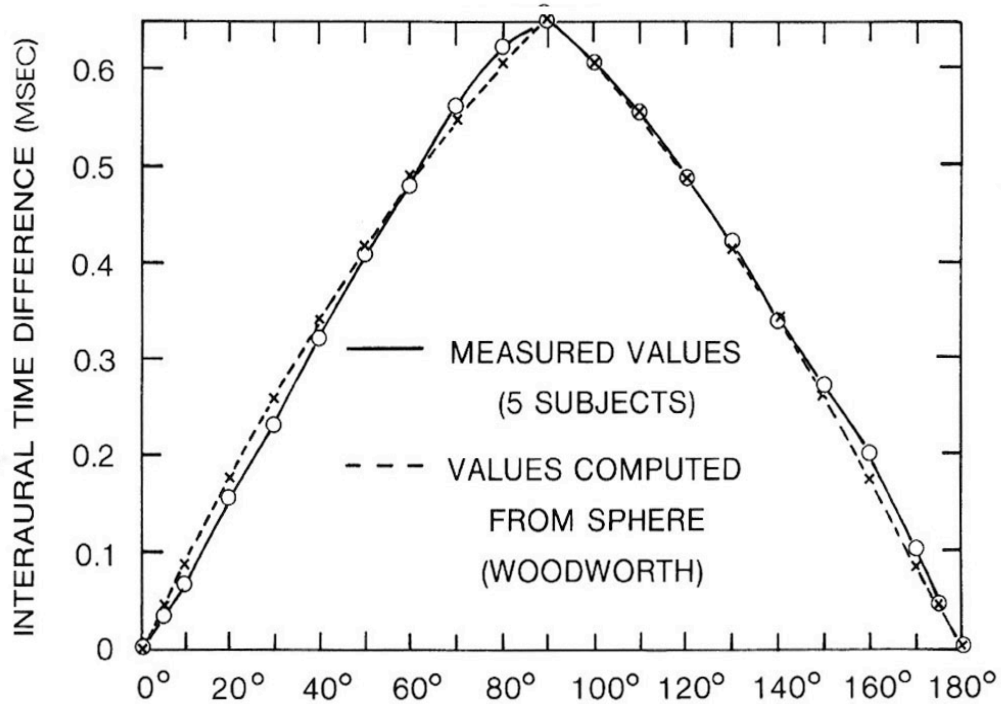
Lateral spatial image shift

ILD (interaural level difference) caused by head shadow of wavelengths > 1.5 kHz



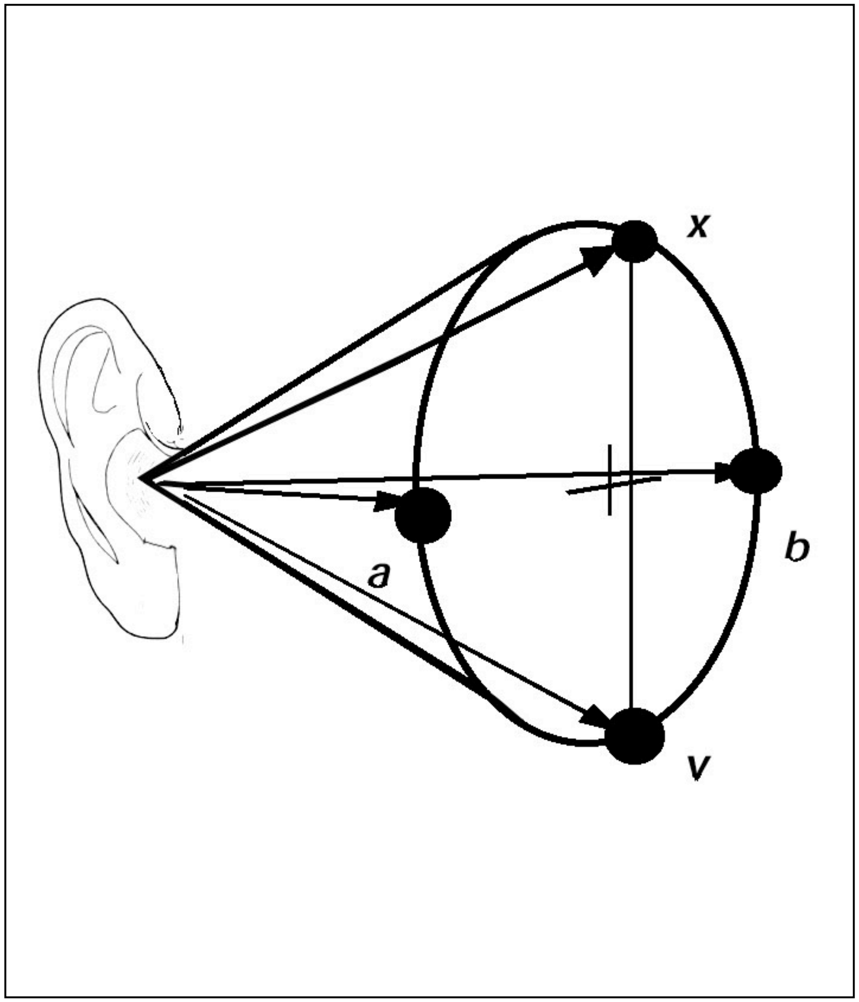
Lateral spatial image shift

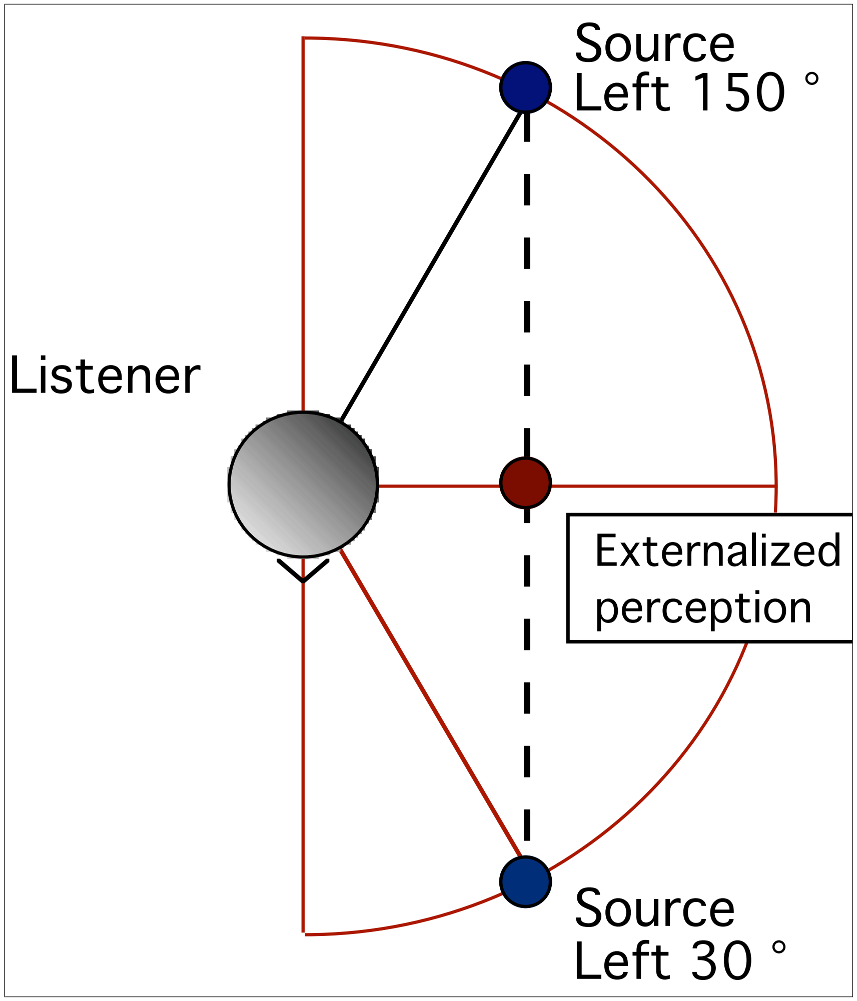
ITD (interaural time difference)



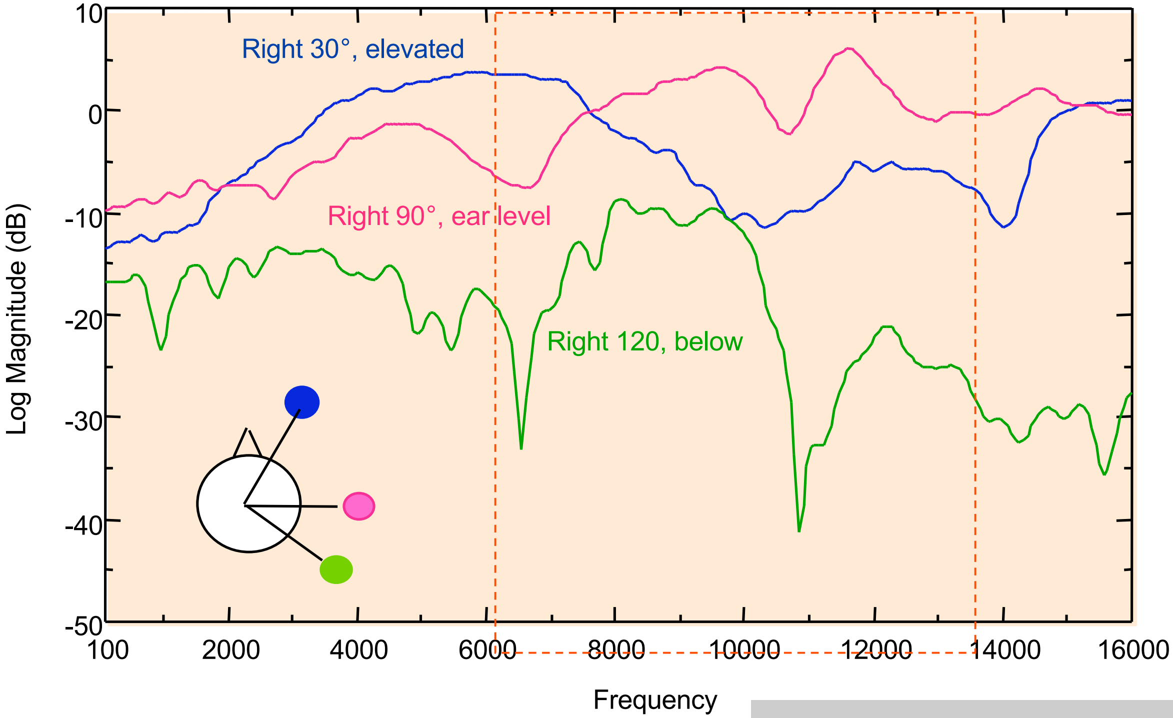
Cone of Confusion Errors

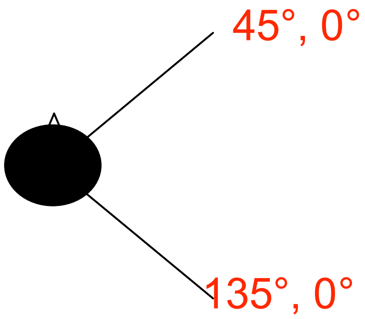
The cone of confusion causes reversals for virtual sources with identical or near-identical ITD or ILD

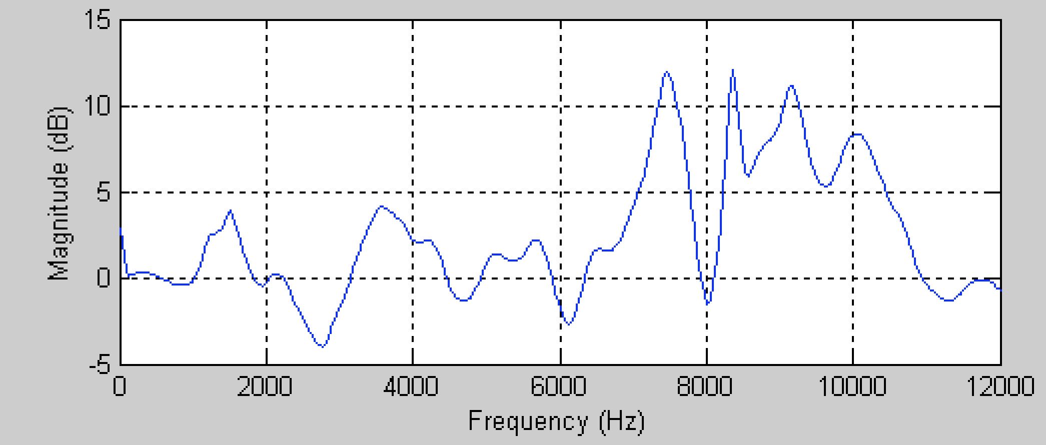




Head-related transfer function cues (HRTFs) provide cues for front-back discrimination and elevation

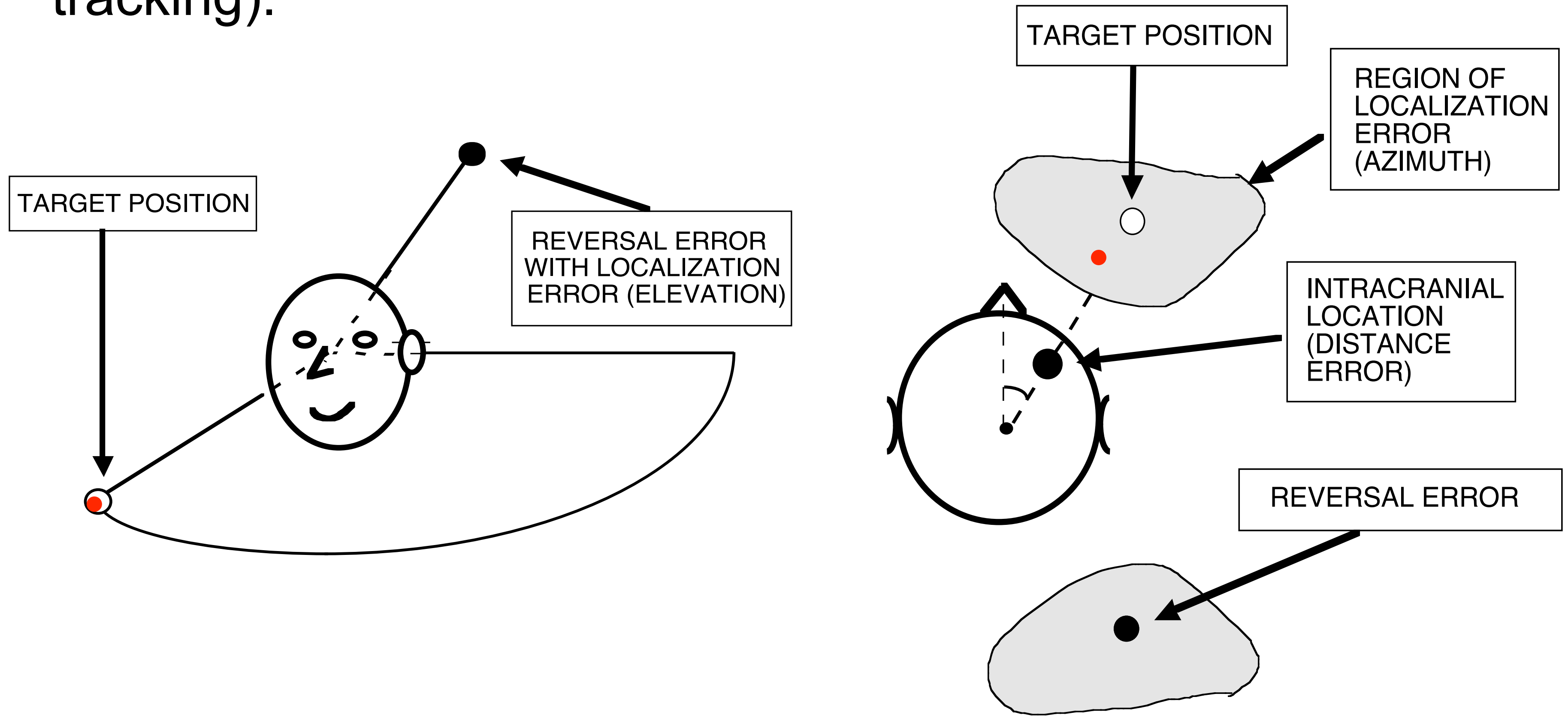




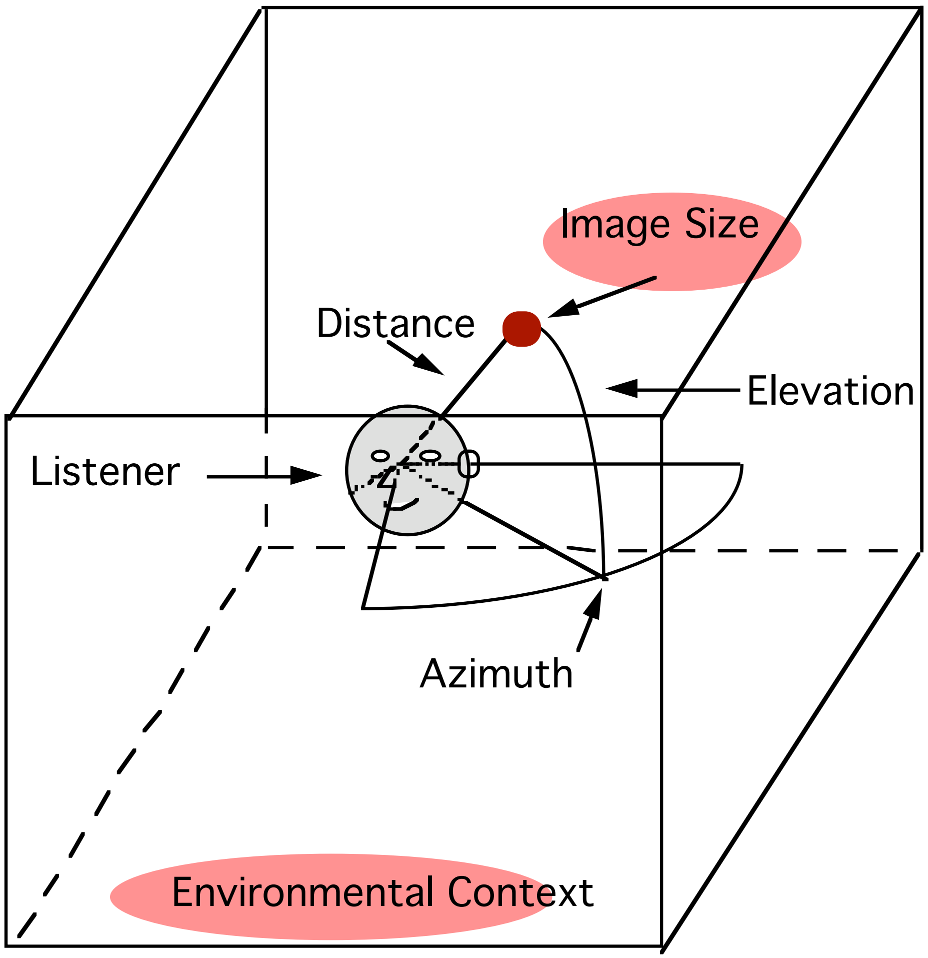


Perceptual Errors

Perceptual errors with headphone-presented spatial sound include **localization blur, inside-the-head localization** (solution: reverberation cues) and **reversals** (solution: head tracking).


TARGET POSITION
REVERSAL ERROR WITH LOCALIZATION ERROR (ELEVATION)
TARGET POSITION
REGION OF LOCALIZATION ERROR (AZIMUTH)
INTRACRANIAL LOCATION (DISTANCE ERROR)
REVERSAL ERROR

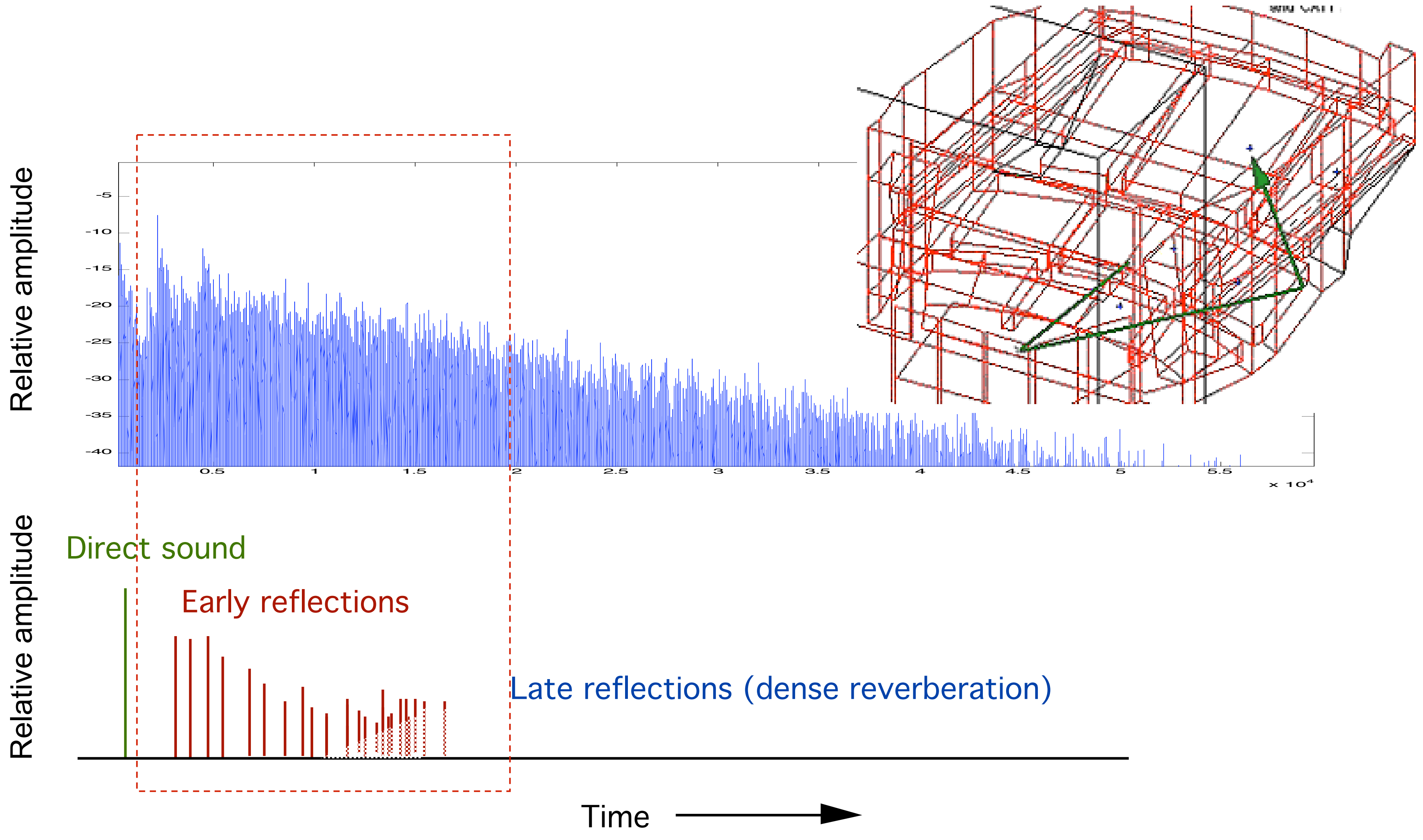
Spatial hearing fundamentally involves perception of the location of a sound source at a point in space (azimuth, elevation, distance).



But a sound source simultaneously reveals information about its **environmental context.**

-reverberation
-image size & extent

Reverberant sound fields provide cues for externalization, distance & environmental context



Direct sound

Early reflections

Late reflections (dense reverberation)

Time

VIRTUAL ACOUSTIC ENVIRONMENTS
Goals / Requirements

Functional equivalence to human auditory system

- Present information accurately in 3 dimensions
- Multiple, simultaneous sources
- Simulation of both static and moving sources
- Real-time, interactive information display
- Head-coupled to achieve a stable acoustic environment
- Modeling of reverberant environments; real-time auralization
- Flexibility in the type of information displayed, e.g., environmental sounds, auditory icons, speech

VIRTUAL ACOUSTIC ENVIRONMENTS
Headphones vs. Loudspeakers

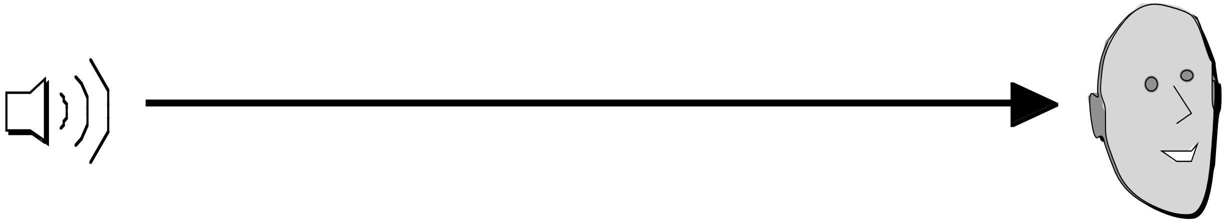
Headphones

- complete control over the acoustic waveforms entering the two ears
- an "infinite" sweet spot: stable acoustic environment even if listener moves around
- full head-coupling of the virtual sources (orientation & position)
- better control of reverberant conditions through modeling
- very low frequency sounds (e.g., explosions) not well synthesized

Loudspeakers

- much harder to control the acoustic waveforms entering the two ears (e.g., need transaural techniques)
- a limited sweet spot: stable acoustic environment only if listener stays relatively still
- limited head-coupling of the virtual sources (orientation only)
- very difficult to control effects of reverberation

Acoustic Scenario Parameters in Headphone-Based Systems



Source

Location
(Implied Velocity)
Orientation
Sound Pressure Level
Waveform
Radiation Pattern
Source Radius

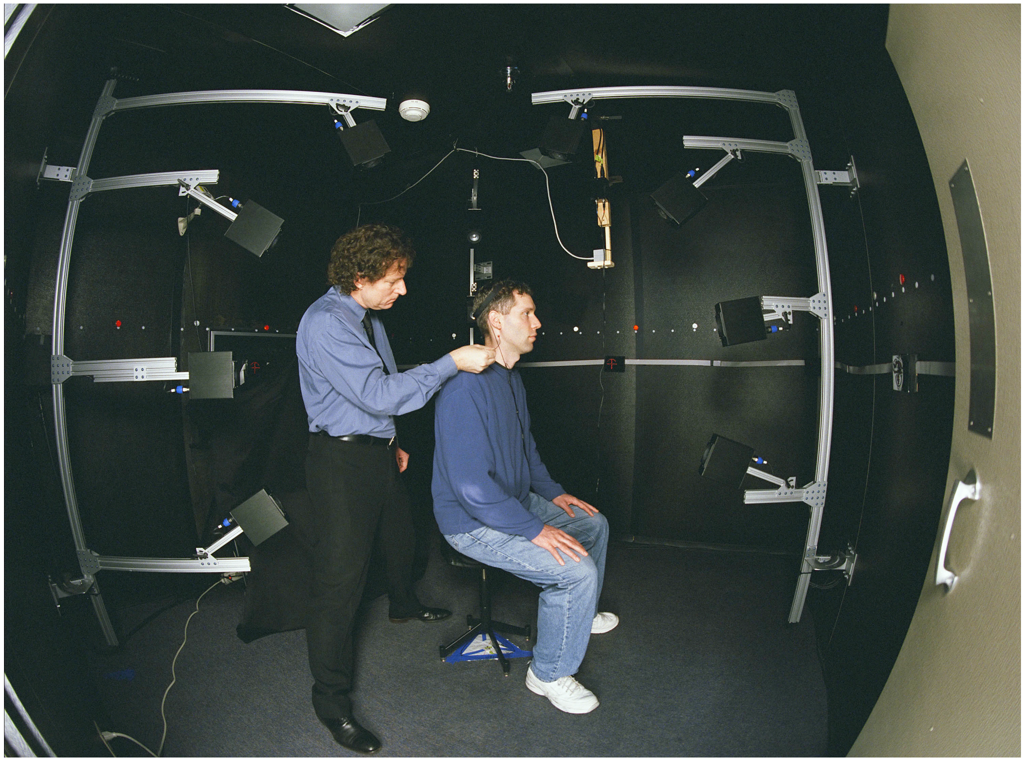
Environment

Speed of Sound
Spreading Loss
Air Absorption
Surface Locations
Surface Boundaries
Surface Reflection
Surface Transmission
Late Reverberation

Listener

Location
(Implied Velocity)
Orientation
HRTF / HRIR
ITD

Measuring HRIRs/HRTFs



Facility at the NASA Ames Spatial Auditory Displays Lab for measuring HRIRs. The 12-speaker system can measure 432 locations at 10 degree intervals in under an hour. The measurement signal is a golay code. Responses are windowed to remove possible reflections.

Psychophysical studies are conducted using headphones to perceptually validate the HRIRs and conduct studies investigating the use of spatial sound in a variety of applications for information display.

Perceptual Research & Engineering Compromises

Perceptual Issues vs. Engineering Compromises

Perceptual	Implementation
Individual differences	Individualized vs. “universal” HRTFs
Localization in reverberant environments	Compromise between localization accuracy & realism, required complexity of models
Distance perception/ externalization	Relative (amplitude) vs. absolute (r/d ratio) distance cueing
Minimum spatial resolution	HRTF measurement techniques, data compression, modeling
Localization Masking	Parameters of simultaneous source display
Head/source motion sensitivity	System latency, HRTF interpolation methods

Perceptual Issues & Implementation Choices

Individual differences in HRTFs (“listening through someone else’s ears”)

- Perceptual impact: increased front/back, up/down reversals, poor elevation perception, poor externalization
- Implementation choices:
  - measure individualized HRTFs
  - model the HRTFs to achieve generalized or adjustable filters
  - head-couple the virtual sources; correlated visual cueing
  - modeling more realistic, reverberant environments
  - develop effective adaptation techniques

Anechoic simulation / inadequate room modeling

- Perceptual impact: poor externalization, inaccurate distance perception
- Implementation choices:
  - relative,dynamic amplitude cues
  - near-field, far-field cues
  - implement more realistic, reflection models (r/d ratio)
  - provide environmental context, room cues/late reverberation

Perceptual Issues & Implementation Choices

Minimum perceptual resolution for spatial location

- Perceptual impact: poor discrimination/spatial resolution
- Implementation choices:
  - develop easier, more accurate HRTF measurement techniques
  - determine required spatial & computational resolution for HRTFs
  - develop perceptually valid interpolation methods
  - model the HRTFs to achieve generalized or adjustable filters

Multiple source presentation

- Perceptual impact: poor resolution/identification of sound sources
- Implementation choices:
  - determine minimum spatial separation
  - determine minimum spectral & temporal differences
  - utilize principles of perceptual segregation, auditory streaming phenomena

Dynamic VAEs: Impact of Latency

Definition of Latency in a Virtual Environment (VE)

End-to-end latency refers to the time elapsed from the transduction of an event or action, such as movement of the head or hand, until the consequences of that action cause the equivalent change in a virtual source/object.

Internal latency refers to a particular system component; e.g., within a spatial audio renderer, it is the delay between acquisition of position data and the rendered audio output.

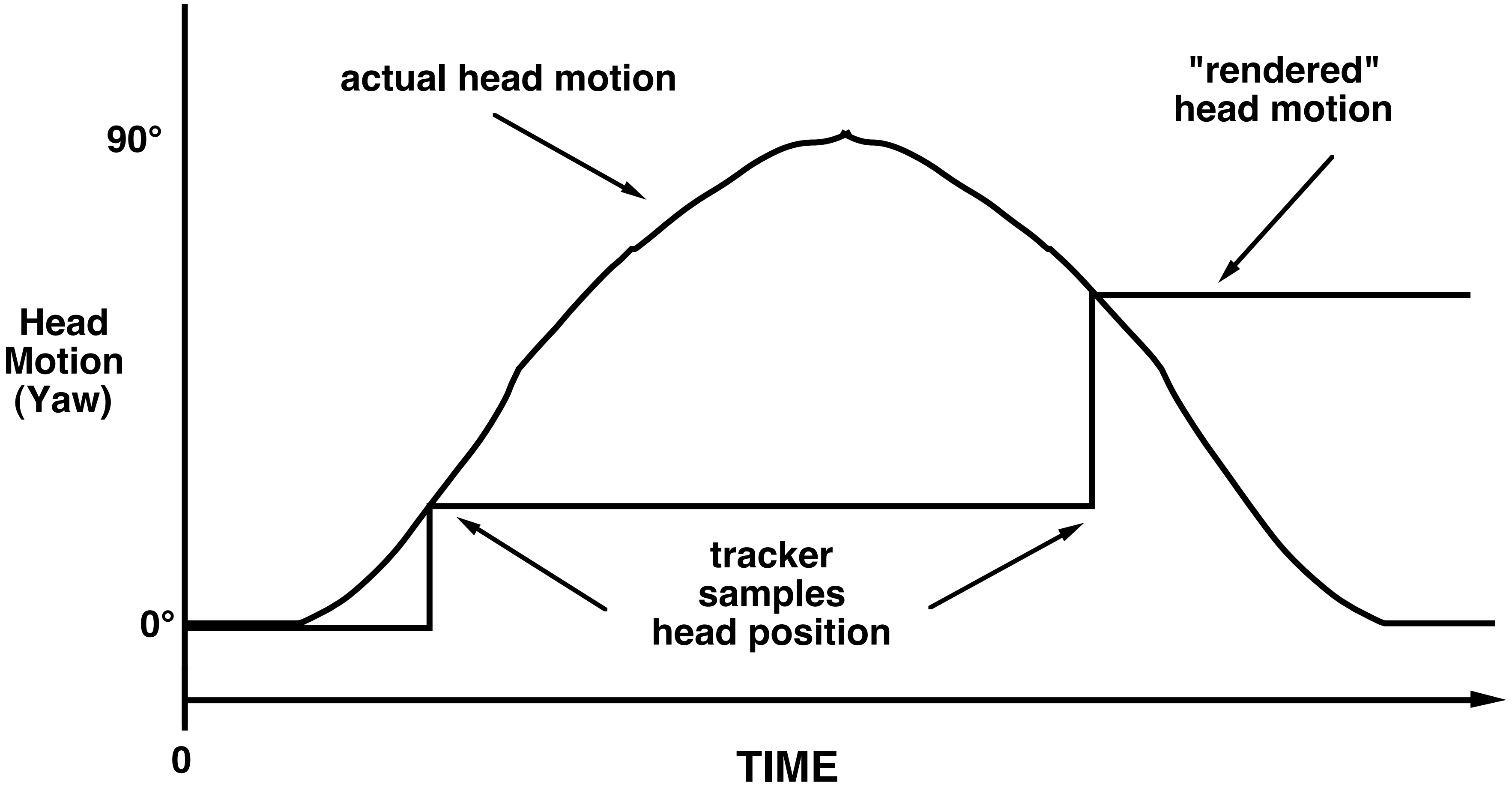
Potential Consequences of Latency

- In any Virtual Environment:
  - Excessive latency degrades the perceived “responsiveness” and a low update rate degrades the apparent “smoothness” of the dynamic interaction

- Motion Compression

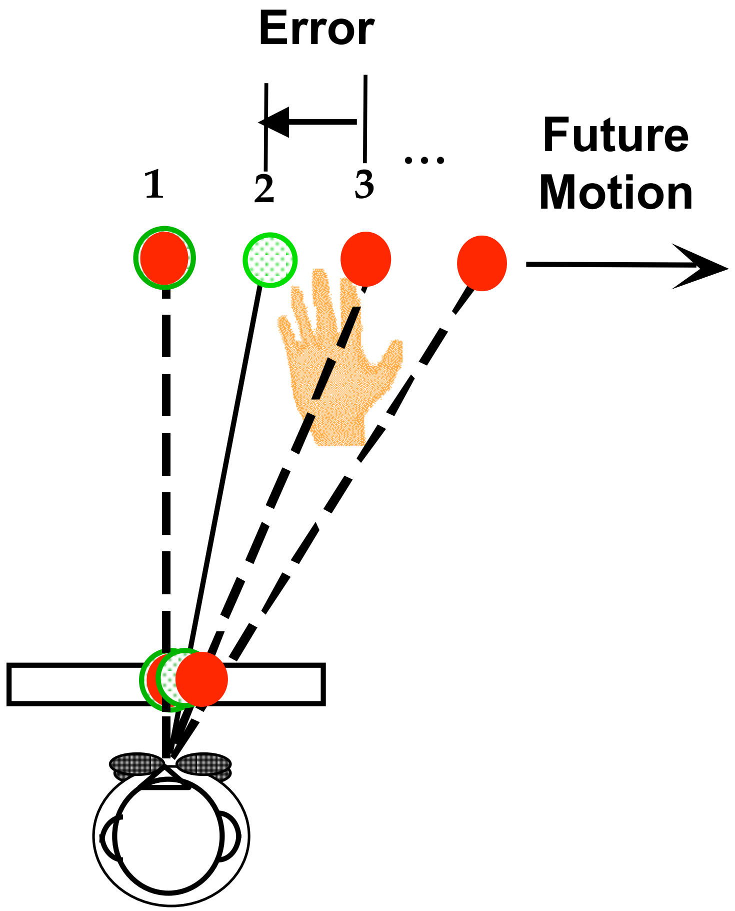
- Positional Instability

Effects of Latency: Motion Compression

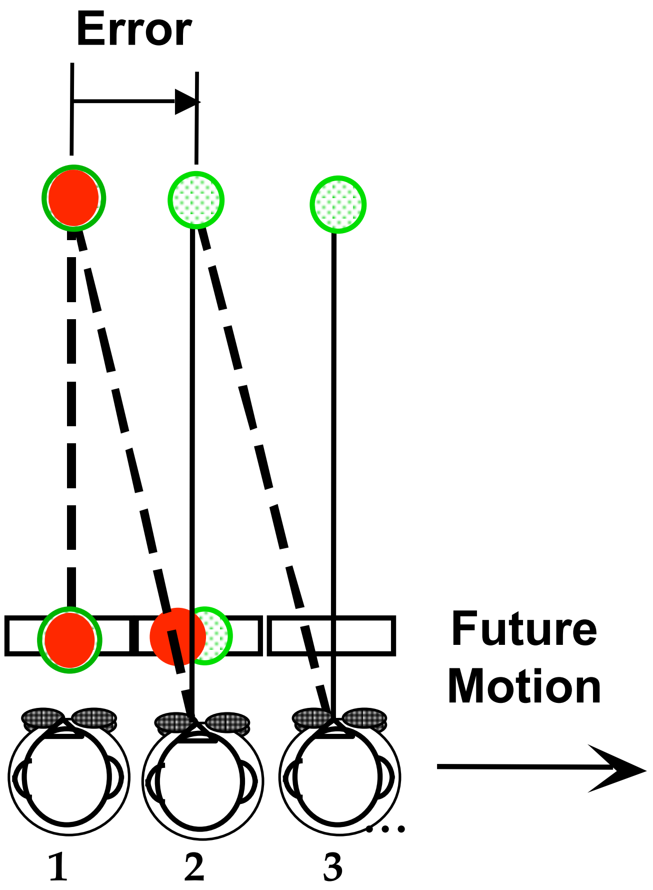


Effects of Latency: Positional Instability

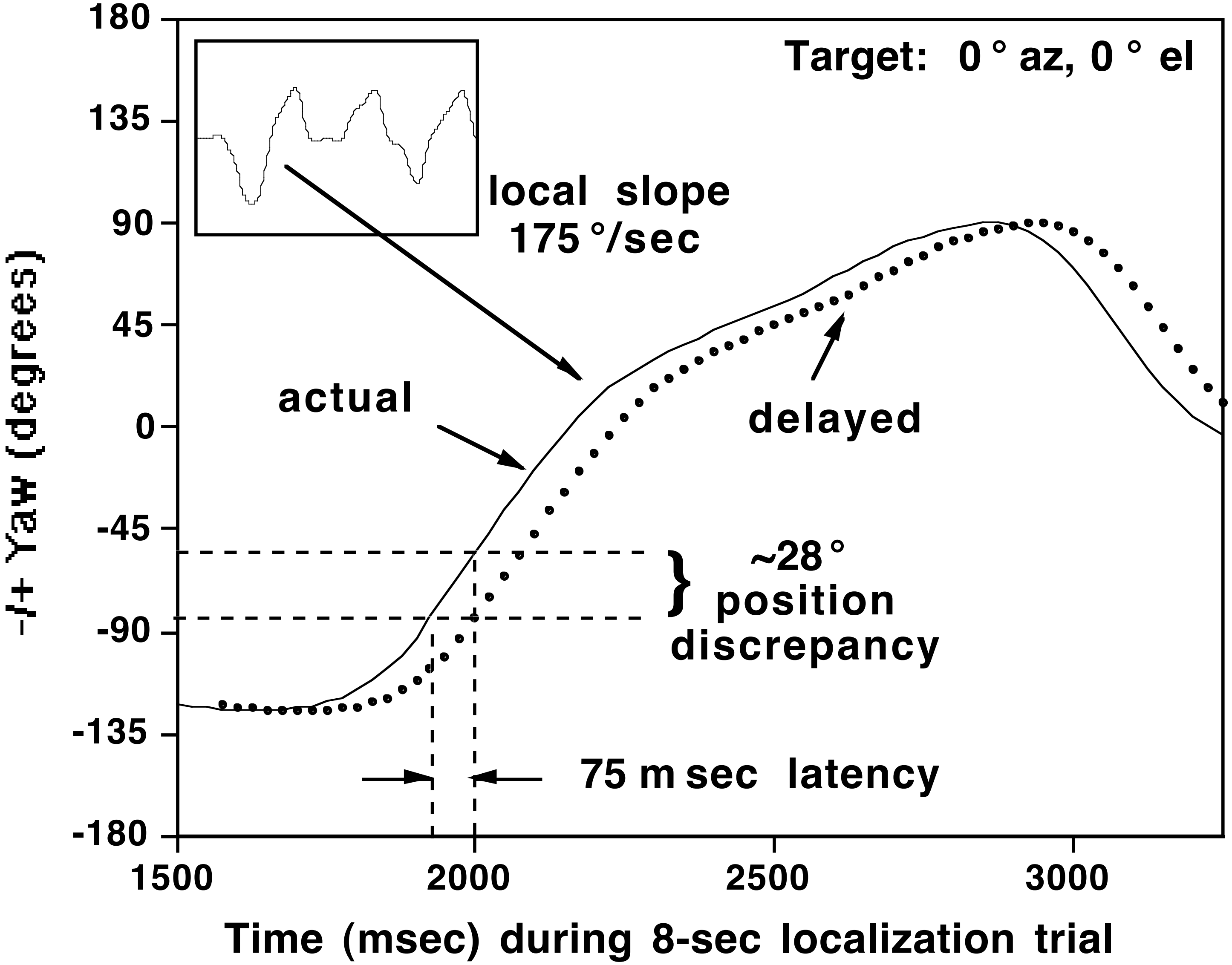
Hand Translation



Head Translation



Effects of Latency: Positional Instability



Importance of Dynamic Simulation in an Auditory VE

- Enabling head motion improves localization, for both individualized & non-individualized HRTFs
  - decreased front-back confusions, ~5% rates comparable to "real world" sound sources;
  - improves externalization somewhat

Perceptual Impact of System Latency

Evidence from Auditory Modality:

From studies of the minimum audible movement angle for real sound sources with listener position fixed (Perrott & Musicant, 1977), one can infer that the minimum perceptible end-to-end latency for a virtual audio system should be about:

source velocities:	90°/s:	92 ms.
	180°/s:	69 ms.
	360°/s:	59 ms.

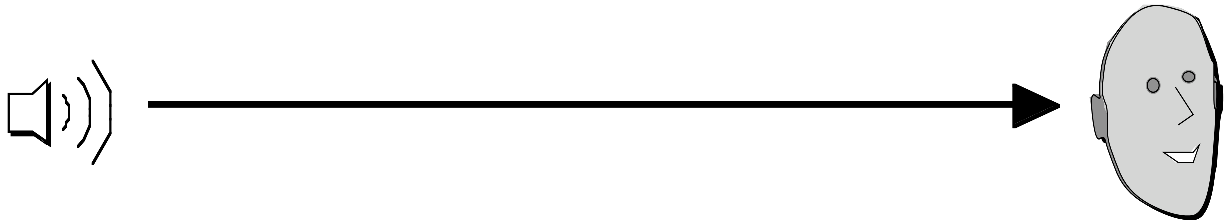
Perceptual Impact of System Latency

Virtual Audio Localization Task with Head-Motion Enabled

- Sandvad (1996):
  - Latencies ≥ 96 ms increased the standard error of localization judgments & increased the elapsed time (< ~3 s) to complete the task.
  - Means of judged locations & confusion rates not reported.
- Wenzel (2001):
  - For longer stimuli (8 s), latency must be as large as 500 ms to reduce localization accuracy & the effect is not large.
  - For shorter stimuli (3 s), latency was somewhat more disruptive; front-back confusions were significantly greater with a 500 ms latency.
  - Latency was not readily noticed until it reached 250 ms.

Comparison of VAE Systems

Acoustic Scenario Parameters in Headphone-Based Systems



Source

Location

(Implied Velocity)

Orientation

Sound Pressure Level

Waveform

Radiation Pattern

Source Radius

Environment

Speed of Sound

Spreading Loss

Air Absorption

Surface Locations

Surface Boundaries

Surface Reflection

Surface Transmission

Late Reverberation

Listener

Location

(Implied Velocity)

Orientation

HRTF / HRIR

ITD

Classes of VAE Systems

- High-fidelity systems for applications like psychoacoustic research, information display, auralization of rooms, & virtual reality
  - Information display: Requires accurate rendering of at least the direct path and early reflections via ray-based models like the image model
  - Auralization: Requires accurate synthesis of the entire binaural room response
  - Virtual presence, realism and directional accuracy all require head-tracking with corresponding attention to system dynamics (latency, update rate)

- Lower-fidelity systems for applications like games & entertainment
  - Rendering algorithms are proprietary
  - Appear to emphasize efficient late reverberation modeling
  - Not clear if the direct paths or early reflections are modeled independently and/or updated dynamically

Trends in VAE Systems

- “Hybrid” High-Fidelity VAE Systems
  - Provide real-time processing of the direct path and early reflections with good system dynamics (low latency: < 70 to 100 ms,
    high update rate: > 10 Hz at minimum)
  - Most employ the image model using convolution in the time domain for real time processing
  - Mix in later reflections / late reverberation that may be updated less frequently or, more typically, remain static with listener motion
  - Later reflections and late reverberation may be derived from:
    - measured or pre-computed Binaural Room Impulse Responses stored in a large database
    - artificial reverberation algorithms whose parameters are based on room acoustic simulations
  - Some enable loudspeaker presentation (cross-talk cancellation / VBAP)
  - Utilize special purpose hardware or distributed computers/CPUs

Trends in VAE Systems

- Software-based VAE Systems
  - Developed for general purpose platforms, such as Intel computers using Windows or Linux
  - Mitigates problems with systems becoming obsolete (discontinued hardware, changing APIs)
  - Processing power, and therefore the complexity of spatial rendering, scales to CPU resources
    - Single CPU enables real-time rendering of several sources and first-order reflections
    - More complex, dynamic room modeling requires multiple CPUs or distributed computer systems
    - Enables an “automatic upgrade” as CPUs get more powerful

Comparison of VAE Systems

VAE System	Audio Display	User Interface	OS	Implementation	Rendering Domain / Room Model
SLAB	headphone	C++	Windows 98/2k	software / Intel	time / image model
DIVA	headphone, speakers	C++	UNIX, Linux	software / SGI	time / image model
AuSIM	headphone	C	client-server (client: Win98/2k, DOS, Mac, etc.)	software / Intel	time / direct path
Spat (IRCAM)	headphone, speakers	Graphical (Max, jMax)	Mac, Linux, IRIX	software / Mac, Intel, SGI	time / reverb engine
AM3D	headphone, speakers	C++	Windows 98/2k	software / Intel (MMX)	? / direct path
Tucker-Davis	headphone	Graphical / ActiveX	Windows 98/2k	special purpose DSP hardware (RP2.1)	time / direct path, reverb engine
Lake	headphone, speakers	C++	Windows NT	special purpose DSP hardware	frequency / BRIR
Convolvotron	headphone	C	DOS	special purpose DSP hardware	time / direct path

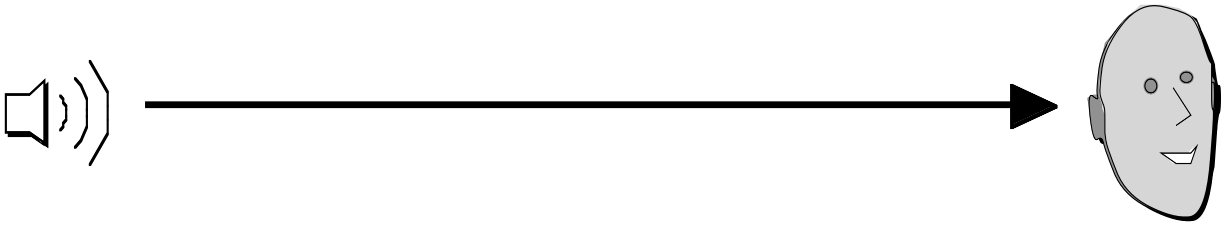
Comparison of VAE Systems (cont.)

VAE System	# Sources	Filter Order	Room Effect	Scenario Update Rate	Internal Latency	Sampling Rate
SLAB	arbitrary, CPU-limited	arbitrary (max. direct: 128, reflections: 32)	image model 6 1st order reflections	arbitrary (120 Hz typical, 690 Hz max.)	24 ms DirectSound (adjustable) 8 ms, ASIO	44.1 kHz
DIVA	arbitrary, CPU-limited	arbitrary, modeled HRIRs (typ. direct: 30, reflections: 10)	image model 2nd order reflections, late reverb	20 Hz	110-160 ms	arbitrary (32 kHz typical)
AuSIM	32 per CPU GHz	arbitrary (128 typical, 256 max.)	N/A	arbitrary (375 Hz default max.)	8 ms (adjustable)	44.1 kHz 48 kHz 96 kHz
AM3D	32-140, CPU-limited	?	N/A	22 Hz	45 ms min.	22 kHz
Lake	1 (4 DSPs)	2058 to 27988	precomputed response	?	0.02 ms min.	48 kHz
Convolvotron	4	256	N/A	33 Hz	32 ms	50 kHz

NASA’s SLAB VAE System

- Software-only solution
  - Scales to CPU resources
  - Anti-obsolescent (discontinued hardware, changing proprietary APIs)
- Object-oriented design - modularity, extensibility, flexibility, maintainability
- Microsoft Windows 2000
  - Developer resources
  - Persistent APIs
  - Low-latency output (DirectSound or ASIO)
  - Price/performance ratio of the Windows/Intel platform
- Written in C++ using the Win32 SDK and MFC

Acoustic Scenario Parameters Modeled in SLAB



Source

Location
(Implied Velocity)
Orientation
Sound Pressure Level
Waveform
Radiation Pattern
Source Radius

Environment

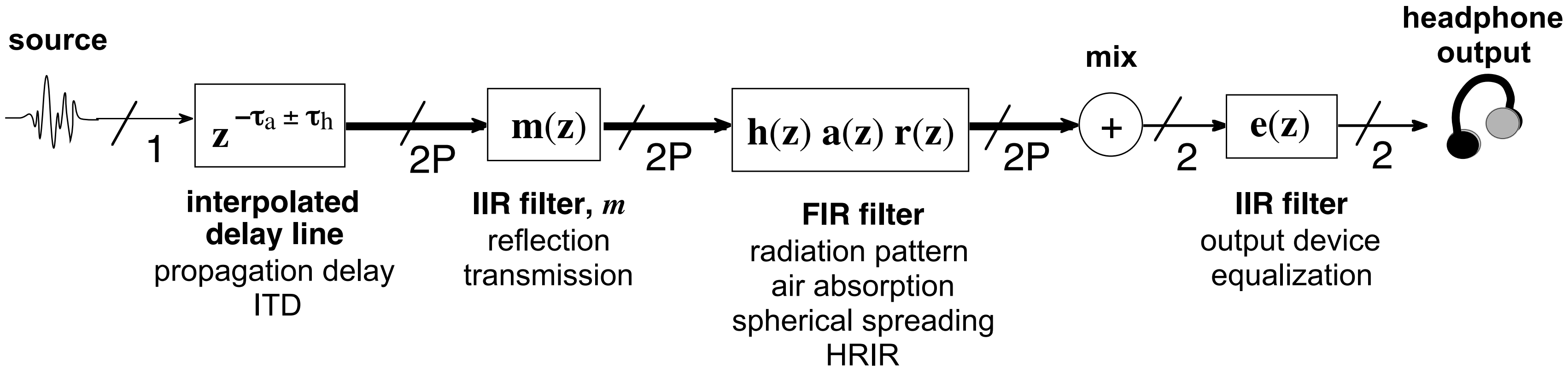
Speed of Sound
Spreading Loss
Air Absorption
Surface Locations
Surface Boundaries
Surface Reflection
Surface Transmission
Late Reverberation

Listener

Location
(Implied Velocity)
Orientation
HRIR
ITD

Spatialization Unit

SLAB Signal Flow



source

$z^{-\tau_a \pm \tau_h}$

interpolated delay line
propagation delay
ITD

$m(z)$

IIR filter, *m*
reflection
transmission

$h(z)\ a(z)\ r(z)$

FIR filter
radiation pattern
air absorption
spherical spreading
HRIR

mix

$+$

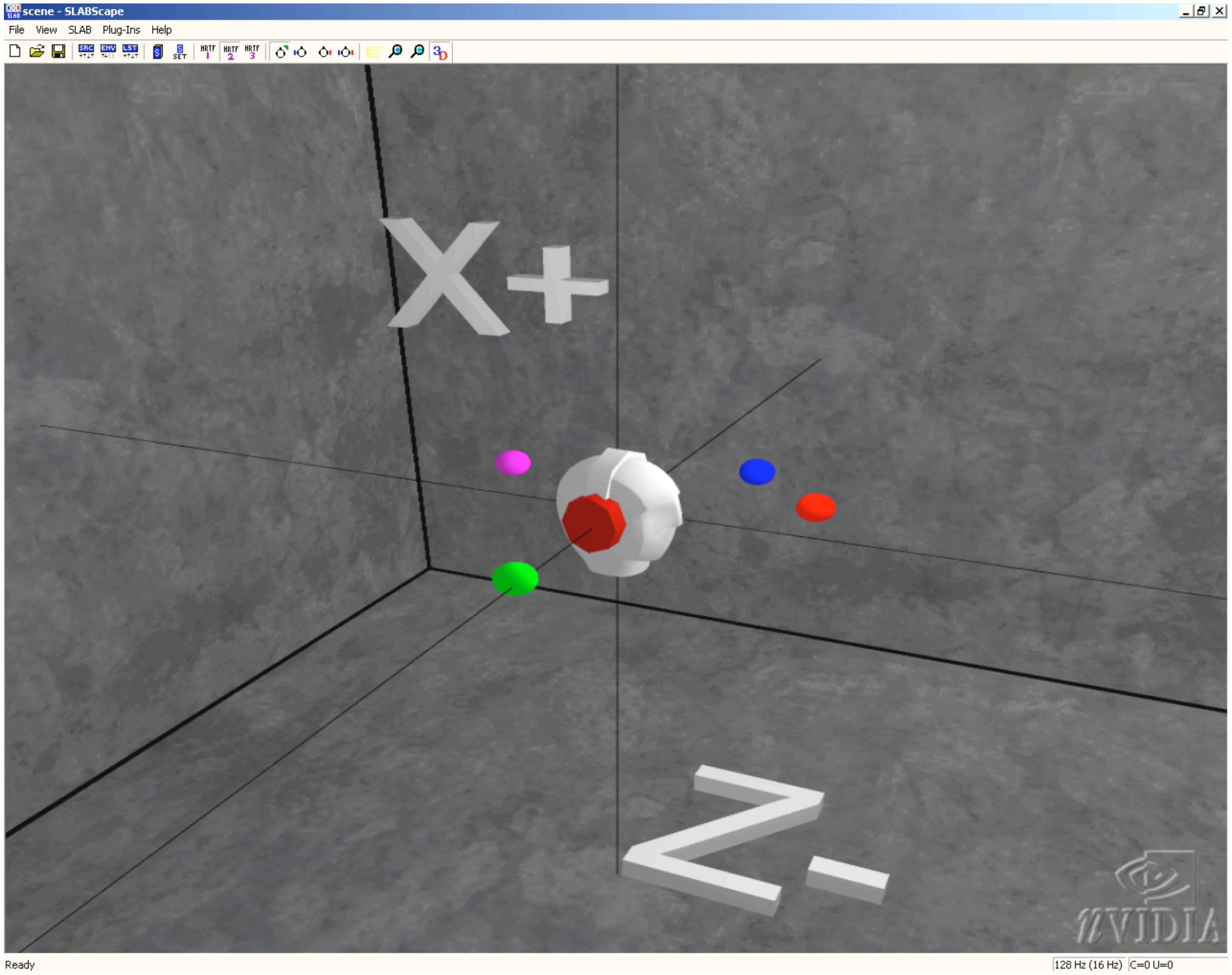
$e(z)$

IIR filter
output device
equalization

headphone output

P = Number of Paths (Direct Path & Reflections); 2P = Paths Rendered for Left & Right Ears

SLABScape Graphical User Interface



SLABScape allows the user to experiment with the SLAB API and access and manipulate the acoustic scenario parameters

Current SLAB Implementation

Signal processing components currently implemented

- interpolated delay line: propagation delay, ITD
- FIR filter: spherical spreading loss & HRIR
- IIR filter: wall materials

Currently not implemented: source radiation pattern, air absorption, surface transmission, and late reverberation

SLAB utilizes a parameter tracking method to smoothly approximate dynamic, time-varying characteristics

Issues in Rendering Dynamic Simulations

Available signal processing structures are static

- Time-varying properties must be approximated
- Need to determine acceptable update rates & latencies to minimize computational & perceptual artifacts
- Perceptually acceptable update rates may be different for different signal processing parameters (HRIRs vs. ITDs & propagation delays)

HRTF database must be sufficiently dense and/or appropriately interpolated to avoid perceptual artifacts

Issues in Rendering Dynamic Simulations

Methods for approximating time-varying properties

- Output cross-fade: Input signal is processed with both past & current rendering parameters and then the two outputs are cross-faded

  * computationally inefficient

  * blend of two different systems introduces artifacts

  * overlap-add methods in frequency domain effectively a form of output cross-fade

- Parameter cross-fade: A set of rendering parameters is cross-faded before processing of the input signal

  * computationally more efficient

  * minimizes artifacts for intermediate states

Rendering Dynamic Simulations

Approximating time-varying properties in SLAB

Parameter tracking: A form of parameter cross-fade that smooths the transition between intermediate states

* computationally more efficient than output crossfade
* minimizes artifacts for intermediate states
* minimizes the impact of noise, e.g., head-tracking error
* SLAB uses a "leaky integrator" to perform a nonlinear cross-fade; computes a running average of signal processing parameters, with greater weight given to the most recent parameters as it tracks toward the desired parameters
* trades smoothness for some additional latency determined by the time constant of the integrator (typically 15 ms in SLAB)

SLAB Dynamic Performance

SLAB dynamic performance characteristics as good or better than other virtual audio systems

Scenario update rate:	120 Hz (8.33 ms)
ITD update rate:	44.1 kHz (22.7 µs)
FIR coefficients update:	690 Hz (1.45 ms)
Internal system latency:	24 ms (DirectSound 8.1 SDK)
	6 ms (ASIO 2.0 SDK)

Scenario Specifications
Rectangular room: Image model
Number of first-order reflections: 6
Number of direct path FIR taps: 128
Number of reflection FIR taps: 32

SLAB Future Directions

- Recent progress: an ASIO-based version of SLAB (v 5.4.1) was developed that enables multiple, full-duplex input/output channels with 6 ms internal latency. ASIO (Audio Stream Input/Output) is a low-latency sound device driver/architecture.
- Future goals include:
  - Source radiation pattern
  - Air absorption
  - Surface transmission
  - Late reverberation
  - Complex room geometries
  - Higher order reflections
  - Multiple processor systems / distributed architecture
- SLAB can be downloaded for free for non-commercial use at http://human-factors.arc.nasa.gov/SLAB/

Conclusions

- VAEs can provide behavioral performance advantages like enhanced situational awareness & multiple-channel presentation for a wide variety of applications.
- Although many of the basic cues needed for spatial sound synthesis are well-understood, there is still a need for continued perceptual research
- Current VAEs tend to fall into two classes, depending on their application goals; e.g., high-fidelity systems for research & auralization and lower fidelity systems for entertainment.
- High-fidelity systems often utilize a “hybrid” approach to spatial rendering involving dynamic, real-time processing of the direct path & early reflections combined with static late reverberation.
- There is an increasing trend toward software-only rendering solutions.
- While many systems still emphasize headphone listening, there is increasing interest in multi-channel loudspeaker systems.
- All VAE technology development can benefit by research which examines the acoustic parameters needed for accurate perception and provides guidance about how best to devote computational resources